Data Science in the Business Environment: Skills Analytics for Curriculum Development

Jing Lu

University of Winchester, Winchester UK, SO22 5HT
Jing.Lu@winchester.ac.uk

Abstract. Data science is an interdisciplinary field of methods, processes, algorithms and systems to extract knowledge or insights from data. University of Winchester Business School, UK is developing an undergraduate degree programme in Data Science which brings together student-centred and business-driven approaches: positioning the course for the interests of students and requirements of employers. The new programme follows the expectations of relevant subject benchmark statements and is built on activities which focus on different aspects of data science, drawing on some existing modules as a base. It integrates key themes in information management, data mining, machine learning and business intelligence. This paper presents the ongoing development of the Data Science programme through the key aspects in its conception and design. Understanding the employment market while defining specific skills sets associated with potential graduates is always important for courses in higher education. The Skills Framework for the Information Age (SFIA) has been adopted and a novel mapping proposed for the interpretation of employability skills related to data science. These are then linked to an adapted process model as well as the specialist modules across academic levels.

Keywords: Subject Benchmarks, Skills Frameworks, Business Analytics, Data Mining, Machine Learning, Business Intelligence, Analytical Tools, SFIA.

1 Introduction

Data Science is an emerging field that requires multi-disciplinary principles to guide the extraction of knowledge from data. In the Business context, the ultimate goal of data science is improving decision making and its links to Big Data and other data-driven technologies. Within the University of Winchester (UoW) Business School in the UK, a new BSc (Hons) Data Science programme is under development which recognises the increasing importance to organisations of knowledge as a commodity. The curriculum is adopting a distinctive structure and pedagogy, building on the well-established Digital & Technology Solutions Degree Apprenticeship as well as the newly-validated Computer Science suite of courses. This is articulated particularly through some specialist modules where technology and business-oriented activities are designed to focus on different aspects of data science, namely: information management, data mining, machine learning and business intelligence.

This paper describes the ongoing development of the Data Science programme and starts with the expectations of relevant QAA subject benchmark statements in the UK, including Business and Management; Computing; and Mathematics, Statistics and Operational Research. Understanding the employment market while defining specific skills sets associated with data science is important for corresponding courses in higher education. Within the same section 2, the national Skills Framework for the Information Age (SFIA) has been adopted to provide the necessary underpinning for the programme, which allows a novel interpretation of data science skills.

Section 3 extends the theme of data science in practice by adapting a cross-industry standard process model as a methodology to guide relevant activities and tasks, which are linked to SFIA-related skills from a business-driven perspective. Analytical tools and publicly available data sources have been recommended here in order to facilitate student projects in terms of data pre-processing, visualisation and analytics.

The Data Science curriculum design has been illustrated in section 4 through a graphical representation across the three academic levels, which gives an indication of the specialist modules versus the more diverse. All of the specialist modules are then linked with relevant SFIA skills through a visual mapping. The paper draws to a close with some concluding remarks and a pointer to future work in relation to the EDISON Data Science Framework.

2 Academic and Professional Frameworks

2.1 Subject Benchmark Statements

Considering Part A of the UK Quality Code for Higher Education, which covers *setting and maintaining academic standards*, there is a range of Subject Benchmark Statements which UK universities are required to meet across their undergraduate provision [9]. There is no particular statement for Data Science as yet, but it is relevant to consider three current subject benchmarks in the context of this paper.

The Business and Management benchmark statement from 2015 [10] generally applies to the various honours degree courses in business studies and management studies, including (e.g.) organisational development and strategic management. However, it can also be used to inform a wider provision, including those courses focused on business functions or sectors. A broad, analytical and highly integrated study of business and management is expected within a framework encompassing organisations, business environment and management. Environment here comprises a range of factors, notably the digital and technological, while management includes rational analysis and other processes of decision making within organisations.

Graduates from Business Schools should be able to demonstrate knowledge and understanding in several areas: one of these is information systems and business intelligence. Skills of particular relevance include problem solving and critical analysis; research – ability to analyse and evaluate a range of business data, sources of information and appropriate methodologies … for evidence-based decision making; and numeracy – use of quantitative skills to manipulate data, evaluate, estimate and model business problems [10].

The next subject benchmark considered here is that for Mathematics, Statistics and Operational Research from 2015 [12]. It applies to cognate programmes of study in MSOR including (e.g.) computational mathematics, numerical analysis and statistical modelling. There are so many real-world applications of mathematics, which has its roots in the systematic development of methods to solve practical problems in areas such as construction and commerce. Understanding of the world is facilitated by identifying and codifying patterns, enabling deeper relationships to be found than could otherwise have been possible from observation or unaided reasoning.

Statistics has been characterised in the MSOR statement as the science of drawing conclusions from data. It includes methods for describing and visualising data to reveal patterns within it as well as the underlying processes producing such data, to extract information and predict future outcomes. The subject area of analytics has become increasingly associated with operational research in recent years. While the name OR is generally well understood, some provision has adopted other titles across the sector, notably: management science, business analytics, business decision methods and business systems modelling [12].

It is worth noting that the MSOR subject benchmark advises that its statement is unlikely to apply to teaching outside cognate departments, although it is important that such programmes pay due attention to the place of MSOR within them.

Moving on finally to the Computing benchmark statement from 2016 [11], a range of provision across computer science and information systems is addressed. For the purposes of this paper, courses in data management, information modelling, machine learning and knowledge representation are especially relevant. Computing overlaps with a number of adjacent subjects, including (e.g.) mathematics and business. Information systems in particular is concerned with the modelling, codification and storage of data for subsequent analysis – specific areas of interest again relate to databases and information modelling as well as the interactions between information systems and the more socio-technical systems. Those courses focused on Computing in society fall under this subject benchmark if their content is informed by computer engineering, software engineering, information technology or information systems.

Computing-related cognitive skills include an understanding of scientific method and its application to problem solving; knowledge and understanding of modelling for the purpose of (e.g.) prediction; as well as the deployment of methods and tools for implementation of systems. On the other hand, Computing-related practical skills comprise a range of abilities, notably deploying tools effectively in the solution of real-world applications; and critically evaluating and analysing complex problems, including those with incomplete information. Universities are also required to provide every student the opportunity to acquire more generic skills to enhance employability. They include intellectual skills, self-management, team working and significantly *contextual* awareness here, to understand and meet the needs of (e.g.) business and the community [11].

It can be seen that there are elements of all three benchmarks which are relevant to some extent to Data Science and these are developed further in the following sections. In particular it is clear that, if only one subject benchmark statement was selected, it would be that for Computing.

2.2 Professional Skills Frameworks

The Skills Framework for the Information Age (SFIA) describes the skills expected of professionals in roles involving information and communications technology. It has become the globally accepted common language for skills and competencies required in the digital world [13]. SFIA gives employers a framework which they can use to measure the skills they have against the skills they need, and tells education and training providers what the job market wants. It is supported by key organisations such as: BCS (British Computer Society), Tech Partnership (formerly e-skills UK), IET (Institution of Engineering and Technology), IMIS (Institute for the Management of Information Systems) and the IT Service Management Forum (itSMF).

BCS in conjunction with SFIA offer a skills matrix, called SFIAplus [14], which contains the framework of IT skills plus detailed training and development resources. It provides the most established and widely adopted skills, training and development model reflecting current industry needs. SFIAplus can be viewed as a three-dimensional model which comprises Categories of Work – Strategy and Architecture; Change and Transformation; Development and Implementation; Delivery and Operation; Skills and Quality; Relationships and Engagement – as well as Levels of Responsibility and Task Components.

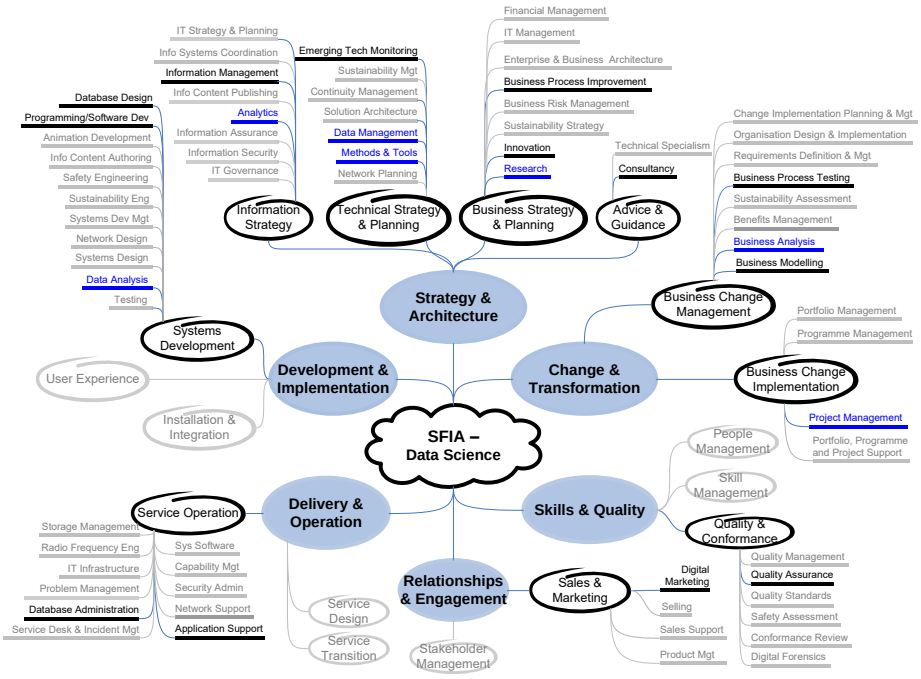

Fig. 1. An interpretation of Data Science skills using SFIA.

The Data Science programme proposed at UoW aims to develop how to use data to provide key business insights, helping companies improve their performance and make

key decisions. Moreover, the programme is designed to meet employers' needs for innovative expertise as well as students' needs for an engaging and developmental course of study leading ultimately to rewarding employment. While the UK Government has published the essential *capabilities* it needs from Data Scientists [3], describing in some detail the knowledge and experience required, the corresponding essential *competencies* comprise those used in the Civil Service and are more generic.

For the Business environment, SFIA has been chosen as the reference against which employability skills are mapped here. A novel presentation of this framework is given in Fig. 1 – individual skills are displayed across six categories of work and associated subcategories – however, the ones which are considered most relevant to data science are shown in bold. It is interesting to note that a key category in this interpretation is Strategy, although Business Change is relatively significant too.

In terms of wider frameworks for data science, the EU-funded EDISON project [2] has focused on activities to establish the new profession of Data Scientist. This has included development of a Data Science Competence Framework (CF-DS) which provides the basis for other components. CF-DS defines five competence groups as Data Analytics; Data Science Engineering; Data Management; Research Methods and Project Management; and Domain-based Business Analytics. Related skills are labelled in blue in Fig. 1 in order to cross reference with SFIA.

3 Methodology and Practice

3.1 Cross-Industry Standard Process Model

There is not an established process model for data science although the most widely used approach for analytics is CRISP-DM, the Cross-Industry Standard Process for Data Mining [15]. Since the data mining process breaks up the overall task of finding patterns from data into a set of well-defined subtasks, it is also useful for structuring discussions about data science. Fig. 2 shows the process model adapted for data science based on activities and tasks linked to SFIA-related skills. At the centre of the model is data management, which may include the internal data environment within an organisation and the external data sources as necessary.

Business Knowledge and Understanding. Prior to the start of a data science project it is crucial to incorporate as much insight as possible into the business goals – then specify business questions and determine any other business requirements. It is also important to define the nature of business success for the project.

Data Understanding. This phase involves accessing the data and exploring it in more detail – this will help to determine its quality prior to the data pre-processing phase. Historical data is often collected for reasons unrelated to the creation of a model, so will need to be considered appropriate to the project.

Data Pre-processing. The data chosen to be included in the analysis may be based on the objectives set at the business understanding stage, the quality of the data determined at the data understanding stage or other practical aspects. Data may be constrained by the analytical technologies used to create the model, e.g. it may be required to be in a different format. Preparation will involve all activities required to

construct the final dataset including selecting attributes, cleaning the data to address any data quality issues and transforming data to create derived variables.

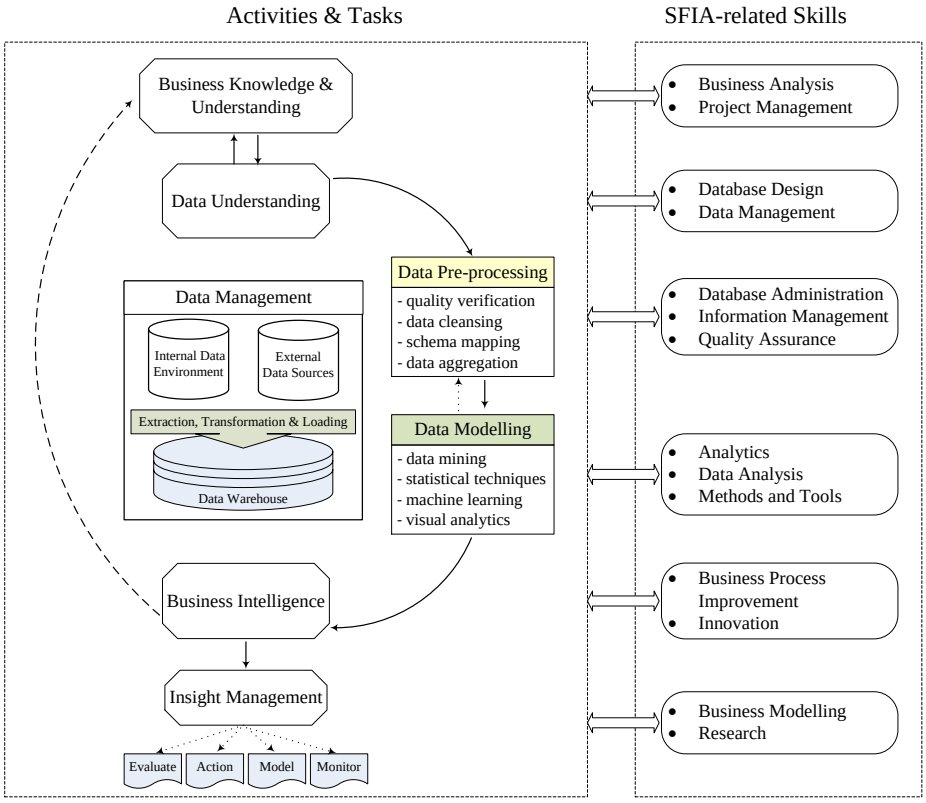

Fig. 2. A process model for Data Science based on CRISP-DM.

Data Modelling. Various modelling approaches will be deployed based on the business objectives and the dataset which is used [5]. Statistical analysis, data mining and machine learning are fundamentally involved with extracting information from a dataset. Common analytical techniques are classification, clustering, regression and dimension reduction while visual analytics technology combines data analysis with data visualisation and human interaction.

Business Intelligence. The primary goal of data science for business is to support decision making – business intelligence focuses on supporting and improving the decision-making process. The modelling results will need to be evaluated carefully as "various stakeholders have interests in the business decision-making that will be accomplished or supported by the resultant models" [8].

Insight Management. The results from the modelling and subsequent evaluation will determine how the model will be deployed to make improvements in organisations.

This could include implementing a predictive model into pre-existing information systems [1]. This stage will also involve planning of the maintenance strategy.

While CRISP-DM has been the industry standard for data mining over the last two decades, Stirrup argues that the model has not been updated to work with new technologies – such as big data – and recommends use of the "Team Data Science Process" cyclical model to address these issues [16].

3.2 Analytical Tools and Data Sources

Data scientists need to be proficient in understanding, searching, extracting and presenting information from structured and unstructured data sources. Keeping up-to-date with the latest trends in technological development is key for effective analytics. Table 1 provides an illustration of some analytical tools associated with the SFIA "Analytics" skill from a technical perspective.

Table 1. Analytical tools and techniques.

SFIA "Analytics" – Typical tools and techniques	Excel	XLMiner	Alteryx	SPSS	R	iNZight	Weka	Tableau	Python
Statistical analysis and forecasting	√	√		√	√	√	√	√	√
Machine learning and data mining		√		√	√		√		√
Graphical visualisation of data	√	√		√	√	√	√	√	√
Data and information modelling			√		√		√		√
Decision support systems		√	√					√	√

As a spreadsheet, Excel can be used for data entry, manipulation and presentation, but it also offers a suite of statistical analysis functions and other tools that can be used to run descriptive statistics and perform inferential statistical tests. In addition, XLMiner is the comprehensive data mining plug-in for Excel, now known as Analytic Solver.

Alteryx is a tool especially made to extract, transform and load data into a data warehouse. Its key capabilities for data preparation include: connect to and cleanse data from data warehouses, spreadsheets and other sources; improve quality of data with profiling, advanced data cleansing and data manipulation tools; repeatable workflow design to assist with data integrity during data preparation process.

SPSS is a software package which has been widely used for statistical analysis by social scientists, education researchers, health researchers, market researchers, survey companies, government and other organisations for many years. IBM SPSS Modeler is a data mining and text analytics software application used to build predictive models and conduct other analytical tasks.

R is a language and environment for statistical computing and graphics, with RStudio providing a user-friendly interface to analyse and manipulate data. R is commonly used for big data management and analysis – it is widely accepted in the data science field and has a very active support community. Developed using R, iNZight can also generate insights into real-world data by producing graphs and summaries through statistical analysis.

Weka (Waikato Environment for Knowledge and Analysis) is open source software written in Java [4] which offers a wide range of statistical inference and machine learning algorithms. It contains tools for data pre-processing, classification, regression, clustering, association rules, sequential patterns mining and visualisation. It provides a way to easily test the performance of a comprehensive suite of data mining and machine learning algorithms on real-world problems.

Tableau Software provides a collection of interactive data visualisation products designed for business intelligence. Its advanced analytics functionalities include: cohort analysis through drag-and-drop segmentation; what-if analysis by modifying calculations and testing different scenarios; and predictive analysis using trending and forecasting models. In addition, an R plug-in allows integration with other platforms.

Python has become an even more popular and powerful programming language in the era of data science. Data analysis, machine learning, information visualisation and text analysis techniques can be applied through Python software libraries and toolkits such as pandas, scikit-learn, matplotlib and nltk to gain further insight into data.

Table 2 is a list of some useful resources for data science projects in the areas of data cleansing, visualisation, data mining and machine learning – the data sources column contains hyperlinks to the individual repositories.

Table 2. Data sources for Data Science projects.

	Data Sources	Description
Data Cleansing	data.world	A social-based data source that allows users to share/clean/ improve data collectively. Can write SQL within the interface to explore data and join multiple datasets.
	The World Bank	The platform provides several tools like Open Data Catalog, world development indices, education indices etc.
	Reddit	A community discussion site which has a section devoted to sharing interesting datasets.
Data Visualisation	FiveThirtyEight	Interactive news and sports site with data-driven articles. Each dataset includes the data, a dictionary and the link to the story.
	FlowingData	Catalogue of data sources, described in detail and shown with examples. It explores how statisticians, data scientists and others use analysis and visualisation.
	Tableau Public	Sample data for visual analytics in the categories of Education, Public, Government, Science, Technology, Health, Business, Sports and Entertainment etc.
Machine Learning	UCI Machine Learning Repository	One of the oldest and most famous sources of datasets online. Vast majority are clean and ready for machine learning.
	Kaggle	A data science community which hosts machine learning competitions – contains externally-contributed datasets.
	Quandl	For financial and economic datasets – useful for building models to predict economic indicators or stock prices.

4 Education and Training

4.1 UG Programme Development

Within the UoW Business School, the BSc Digital & Technology Solutions degree apprenticeship and BSc Computer Science suite both inform the BSc Data Science prototype. One way to express the significance of current modules to the new undergraduate programme is to display the relationships graphically. Fig. 3 shows the extent to which relevant modules may contribute to Data Science – each of the individual boxes represents a module with the colour-coding across Level 4 to 6. The boxes within the triangle are the specialist modules proposed for data science while the oval shapes outside indicate diverse modules from other programmes, with dotted ovals for optional modules. There is also a Group Project module for Level 5 and a double-credit Data Science Project for Level 6.

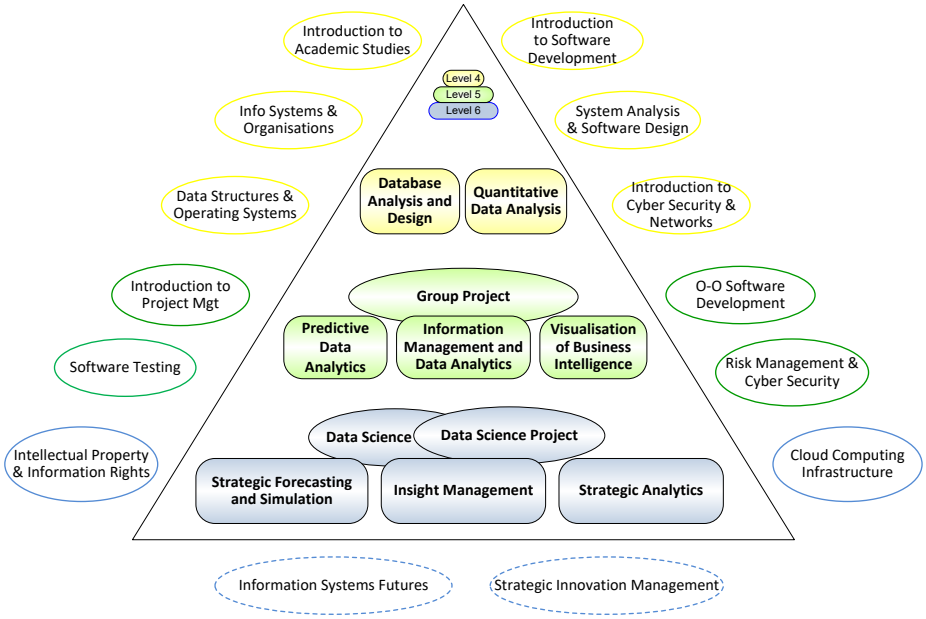

Fig. 3. BSc Data Science prototype.

A brief description for each specialist taught module is given below – these are linked with Data Science Body of Knowledge areas from the EDISON project [2], which are associated with their CF-DS competence groups.

Database Analysis and Design. Introduces analysis and design concepts (using SQL and UML) that are essential for developing and implementing relational database solutions in given business scenarios [DSDM/DMS: Data management systems].

Quantitative Data Analysis. Introduces quantitative analytics concepts, procedures and software tools (Excel and SPSS) for specific data analysis tasks [DSDA/SMDA: Statistical methods for data analysis].

Information Management and Data Analytics. Organised around three themes: Database Management and SQL; Data Warehousing and Information Modelling; Data Mining and Knowledge Discovery [DSDA/DM: Data mining].

Predictive Data Analytics. Provides experience of predictive modelling and analytics across a range of domains, acquiring relevant practical skills (using R and Weka) in data science to create data visualisations and carry out analyses [DSDA/PA: Predictive analytics].

Visualisation of Business Intelligence. Focuses on techniques for data extraction and preparation while analysing data in visual ways (using Alteryx and Tableau) to generate insight for business intelligence and decision making [DSENG/IS: Information systems].

Insight Management. Provides knowledge and skills to identify and evaluate a business issue and/or research problem, effectively analyse data and interpret insights (using iNZight and Tableau) so that they can have an impact at managerial levels of organisations [DSBPM/BA: Business analytics].

Strategic Forecasting and Simulation. Covers the data-driven business prediction topics of forecasting and simulation (using R and XLMiner) to develop advanced models and solutions to real-world problems [DSDA/MODSIM: Computational modelling, simulation and optimisation].

Strategic Analytics. Provides students with a deeper understanding of how data is used by strategic decision makers, covering the analysis of big data (using Python and Weka) as well as data analytics case studies [DSDA/ML: Machine learning].

4.2 SFIA-related Skills Mapping

There are relationships between the proposed Data Science modules and the key SFIA skills too, which are demonstrated in Fig. 4. First dotted lines are used to connect related SFIA skills to each other. For example, relevant SFIA skills for "Analytics" comprise Information Management, Data Analysis, Business Analysis and Business Modelling. Similarly, relevant skills for "Information Management" include (e.g.) Database Design, Data Management, Innovation and Business Process Improvement among others.

The specialist modules for the Data Science programme can then be mapped onto corresponding SFIA skills, where the same colour-coding applies as in Fig. 3. For example, modules linked with the SFIA Analytics skill are Predictive Data Analytics, Strategic Analytics and Strategic Forecasting & Simulation. As another example, the Visualisation of Business Intelligence and Insight Management modules are closely connected with the SFIA Business Analysis and Business Modelling skills. Finally, the Group Project and Data Science Project are linked primarily to the Research skill, although the taught specialist modules will all apply to some degree. Fig. 4 represents a novel aspect of skills analytics in the Data Science context.

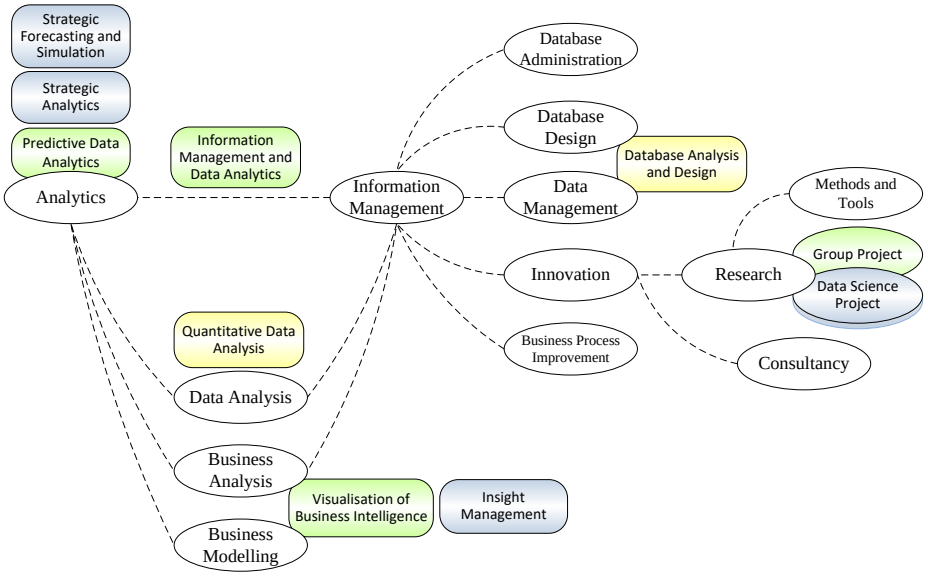

Fig. 4. Linkage between SFIA skills and specialist modules.

5 Conclusion

The digital economy has facilitated an explosion in the data available to the world which has affected businesses, jobs and education. The term "Big Data" refers to datasets so large and complex that it would be impossible to analyse them using traditional methods. Big data was originally defined in terms of the three Vs, namely: "high-*volume*, high-*velocity* and high-*variety*". A 4th V for *veracity* ensued, referring to the trustworthiness of the data. However all this enormous quantity of fast-moving data of different types and confidence levels has to be turned into *value*, which leads to the 5th V for big data [6]. Data Science will help organisations to turn data into valuable insights in order to better understand their customers and optimise their internal processes while identifying cost savings and growth opportunities [7]. Some representative business analytics approaches include for example financial analytics, market analytics, customer analytics, employee analytics and operational analytics alongside the core analytical tools and techniques.

This paper has discussed an overall curriculum design and the skills required for Data Science in the business environment. The new BSc Data Science development is already having a positive impact on other programmes within the University of Winchester Business School, for example: BA Accounting & Finance/Management and their Level 5 Research and Analysis module; MSc Digital Marketing & Analytics and its Analytical Tools for Digital Data module; and the Executive MBA module delivering Insight Management for business professionals.

In terms of the next stage for programme development at Winchester, the EDISON Data Science Framework [2] will be considered further – in particular the detailed Data

Science Model Curriculum. An evaluation of the extent to which their recommended learning outcomes and topics would apply in UK higher education will be significant here, especially within a Business School context. The real evidence of what can be achieved by the programme will begin to materialise following its first year of delivery in 2019/20.

References

1. Abbott, D.: Applied predictive analytics: principles and techniques for the professional data analyst. Wiley, Indianapolis (2014).
2. EDISON: building the data science profession, http://edison-project.eu/, last accessed 6 March 2018.
3. GOV.UK: Data scientist – skills they need, https://www.gov.uk/government/publications/data-scientist-skills-they-need/data-scientist-skills-they-need, last accessed 12 March 2018.
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1): 10-18 (2009).
5. IBM: CRISP-DM help overview, https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm, last accessed 28 February 2018.
6. Marr, B.: Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results, Oxford: John Wiley & Sons (2016).
7. Marr, B.: Key business analytics: The 60+ business analysis tools every manager needs to know. Harlow, Pearson (2016).
8. Provost, F., Fawcett, T.: Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, California (2013).
9. QAA Subject Benchmark Statements for subjects studied at honours degree level, http://www.qaa.ac.uk/assuring-standards-and-quality/the-quality-code/subject-benchmark-statements/honours-degree-subjects, last accessed 9 March 2018.
10. QAA Subject Benchmark Statement: Business and Management, http://www.qaa.ac.uk/publications/information-and-guidance/publication?PubID=2915#.WqKJh2rFLIU, last accessed 9 March 2018.
11. QAA Subject Benchmark Statement: Computing, http://www.qaa.ac.uk/publications/information-and-guidance/publication?PubID=3043#.WqKIk2rFLIU, last accessed 9 March 2018.
12. QAA Subject Benchmark Statement: Mathematics, Statistics and Operational Research, http://www.qaa.ac.uk/publications/information-and-guidance/publication?PubID=2952#.WqKJ4mrFLIU, last accessed 9 March 2018.
13. SFIA: The Skills Framework for the Information Age, http://www.sfia.org.uk, last accessed 28 February 2018.
14. SFIAplus: BCS, http://www.bcs.org/server.php?show=nav.7849, last accessed 28 February 2018.
15. Shearer, C.: The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22 (2000).
16. Stirrup, J.: What's wrong with CRISP-DM, and is there an alternative? https://jenstirrup.com/2017/07/01/whats-wrong-with-crisp-dm-and-is-there-an-alternative/, last accessed 28 February 2018.